MorphoLogic's submission for the WMT 2009 Shared Task

Attila Novák

MorphoLogic

Kardhegy utca 5, Budapest 1116, Hungary

novak@morphologic.hu

Abstract

In this article, we describe the machine translation systems we used to create MorphoLogic's submissions to the WMT09 shared Hungarian to English and English to Hungarian shared translation tasks. We used our rule based MetaMorpho system to generate our primary submission. In addition, we created a hybrid system where the Moses decoder is used to rank translations or assemble partial translations created by MetaMorpho. Our third system was a purely statistical morpheme based system for the Hungarian to English task.

1 Introduction

This year, MorphoLogic submitted translations for the WMT09 shared Hungarian to English and English to Hungarian translation tasks. Our primary submissions were translated by MetaMorpho, a purely rule based machine translation system (Prószéky and Tihanyi, 2002). Since last year's workshop we improved the Hungarian to English grammar of MetaMorpho by making more efficient the handling of certain structural ambiguities and making the way the system handles long sentences more robust.

The way Metamorpho selects the translation to output is not optimal whether or not a full parse for the source sentence could be obtained by its parser.[1] Thus we decided to experiment with a hybrid system where translations and partial translations produced by MetaMorpho are ranked or assembled by the Moses decoder (Koehn et al., 2007) using a target language model.

In addition, we created a purely statistical morpheme based system (also using Moses) for the Hungarian to English task. However, results obtained with the latter setup have been clearly inferior in quality to those produced by the rule based system both in terms of BLEU score and subjective human judgment.

2 The MetaMorpho translation system

MetaMorpho is a rule based system the architecture of which differs from that of most well-known rule based systems: it does not contain a separate transfer component. Its grammar operates with pairs of patterns (context-free rules enriched with features) that consist of one source pattern used during bottom-up parsing and one or more target patterns that are applied during top-down generation of the translation. The architecture of the grammar is completely homogeneous: the same formalism is used to represent general rules of grammar, more-or-less idiomatic phrases and fully lexicalized items, these differ only in the degree of underspecification.

The translation of the parsed structures is already determined during parsing the source language input. The actual generation of the target language representations does not involve any additional transfer operations: target language structures corresponding to substructures of the source language parse tree are combined and the leaves of the resulting tree are interpreted by a morphological generator.

MetaMorpho processes input by first segmenting it into sentences, then tokenizing them and performing morphological analysis on tokens, assigning morphosyntactic attribute vectors to them. This is followed by parsing the network of ambiguous token sequences using the source side of the grammar. Features are used in the grammar to express constraints on the applicability of rules and to store morphosyntactic, valence and lexical information concerning the parsed input.

When no applicable rules remain, translation is generated in a top-down fashion by combining the target structures corresponding to the source

---

[1] In the first case, simply the first translation is output instead of considering all possible translations and selecting the best, while in the second case, the algorithm that combines the partial translations does not check how well the target language side of the pieces fit together.

patterns constituting the source language parse tree. A source language rule may have more than one associated target rule. The selection of the target structure to apply relies on constraints on the actual values of features in the source rule.

Unlike in classical transfer-based systems, word order rearrangement is already determined during parsing the source language input by the applied rules and the values of the features. During generation, the already determined rearranged structures are simply spelled out. The morphosyntactic feature vectors on the terminal level of the generated tree are interpreted by a morphological generator that synthesizes the corresponding target language word forms.

Handling ambiguity is always a difficult problem in a rule based system. MetaMorpho gets rid of alternatives either by using high level heuristics or by specific rules explicitly overriding some more general alternatives. Generally MetaMorpho only generates the first possible translation corresponding to the first parse it produces. In the case of long sentences however, MetaMorpho still may run into the problem of generating too many hypotheses. The solution to this problem originally was simply to abort the parser when it had spent too much time on analyzing a sentence. This resulted in a sequence of words at the end of the sentence remaining untranslated. We managed to alleviate this problem by introducing subsentential segmentation that partitions the input sentence into chunks at presumably safe places (usually clause boundaries).

3 Using a target language model to combine partial parses

During parsing, a hierarchy of partial structures is built by the parser. If the parser fails to produce full parse of the sentence, MetaMorpho reverts to using a heuristic process that constructs an output by combining the output of a selected set of these partial structures covering the whole sentence. These assembled translations are usually suboptimal, because in the absence of a full parse some structural information such as agreement is often lost.

3.1 Pronoun dropping

In the case of Hungarian to English translation, pronoun dropping in Hungarian is a further problem when trying to assemble a translation from partial structures. Since the number and person of the subject and the definiteness of the object (in the case of transitive verbs) is exactly ex-

pressed by Hungarian verbal agreement suffixes, explicit subject and object pronouns may be (and usually are) dropped (unless they are focused or otherwise stressed). The problem is that the same verb forms are used when the subject or object is a full NP. In these cases, however no pronoun is incorporated in the verbal suffix:

Hallja.	*He/she/it hears him/her/it.*
Fred hallja a doktort.	*Fred hears the doctor.*

For single verb forms the MetaMorpho parser only generates English phrases that contain subject pronouns (and in the case of a transitive definite verb like *hallja* also an object pronoun: *he hears it*), because the verb is only represented in the grammar by structures that inherently contain its possible argument structures. This results in extra pronouns appearing in the assembled output translation if there is in fact an overt subject and/or object in the sentence. The same thing applies to 3rd person singular possessive constructions:

háza	*his house*
Fred háza.	*Fred's house.*

3.2 Utilizing the Moses decoder

The original partial structure combination algorithm in MetaMorpho does not utilize a statistical model of the target language. In our experiments, we replaced the original phrase combination algorithm with a statistical model using the Moses decoder hoping that this would improve the translations produced in these cases. We created an interface to the parser that can output all partial parses generated during parsing the input sentence along with their translations.

We directly constructed a phrase table from the partial translations and used the Moses decoder to select the best translation using a surface target language model. We assumed a uniform distribution on the translations in the phrase table (for lack of a better estimation of the translation probabilities) and assigned a zero weight to the phrase model in the Moses configuration. Neither did we use a lexicalized distortion table. The decoder thus selects the best translation based on the language model score assigned to it. In our experiments we used 5-gram language models created from the WMT09 bilingual training data. We could not use language models created from the larger monolingual corpora: the RAM in-

stalled in our test machine was not enough for that.[2]

We experimented with various parameter settings and ways of building the phrase table. While including partial translations in the phrase table for sentences that had a full parse definitely hurt performance, adding all alternative full translations (if the parser managed to parse the whole sentence) to the phrase table and letting the language model select the best one (instead of MetaMorpho defaulting to the first successful parse) improved performance as could be expected. We needed to increase the maximum allowed phrase length parameter from the default to allow the decoder to use the full sentence translations (failing to do so resulted in a serious degradation of performance).

Adding alternative versions of phrases containing possibly spurious pronouns to the phrase table with the pronouns removed or properly modified also had a beneficial effect as this reduced the frequency of extra inserted pronouns in the translations.

While our original phrase assembly algorithm never attempts to reorder the chunks it selects, we did experiment with different distortion parameter settings in the statistical approach since reordering comes for free with the Moses decoder. (Well, there is in fact a price to pay for distortion: a sharp fall in decoding speed.) We found that not penalizing word order changes by the decoder clearly had a detrimental effect on the accuracy of translations. The default distortion limit and penalty (distortion limit was of six words (d=6) in this setting; distortion penalty weight was identical with the language model weight) often resulted in translations with completely out-of-place chunks at the end of the sentence. We got the best results (also in terms of BLEU score) when disallowing distortion altogether even though this results in somewhat disfluent output, especially if the target language is English and the original Hungarian sentence was verb final. Disallowing distortion also made decoding more than ten times faster.

3.3 Results

Unfortunately, even with the best parameter settings that we have found, we managed to achieve only a slight improvement in BLEU scores compared to the original heuristics used in MetaMorpho. The following table lists the (case insensitive) BLEU scores achieved by the original purely rule based system and various versions of the hybrid system on the WMT09 test set.[3]

Hungarian to English	
MetaMorpho	9.96
d=6, no distortion penalty, reassembling full parses	9.62
d=6, distortion penalty, no partial analyses for full parse sentences	9.70
d=0, no distortion, no partial analyses for full parse sentences, pronoun dropping	10.10
English to Hungarian	
MetaMorpho	8.13
d=6, distortion penalty, no partial analyses for full parse sentences	8.22
d=0, no distortion, no partial analyses for full parse sentences	8.44

Although we got slightly better results using the hybrid system, we submitted the output of the original fully rule based MetaMorpho system as our primary submission.

4 A morpheme based Hungarian to English statistical translation system

In addition to the hybrid system above, we also experimented with a statistical system using the Moses toolkit that we used to build a Hungarian to English translation system. The model that we implemented is based on a morpheme based representation of both languages instead of a word form based or factored representation.

4.1 The architecture of the system

The Hungarian side of the WMT09 parallel training corpus was analyzed and stemmed using the *Humor* morphological analyzer (Prószéky and Kis, 1999; Prószéky and Novák, 2005) and we used the *Hunpos* tagger (Halácsy, Kornai and Oravecz, 2007) for disambiguating the morpho-

---

[2] Building lower-order LMs, cutting off singletons, and/or limiting the LM's vocabulary to the most frequent phrases could be possible solutions to that problem as the reviewer of the paper pointed out. We are going to try to solve the memory problem using a combination these techniques in our follow-up experiments.

[3] We first used a cleaned-up version of the WMT08 test set (with typos and badly converted characters fixed) in our experiments. Then we rerun some of the test configurations on the WMT09 test set and got similarly improving results, which we report here.

logical tagging. For English tagging, we used *CRFTagger* (Phan, 2006), a Java-based conditional random fields POS tagger, while stemming was performed by *morpha* (Minnen, Carroll and Pearce, 2001). We used the corresponding *morphg* word form generator to generate the output surface word forms. Unfortunately, morpha neutralizes some present and past forms of the copula, we needed to fix this to get the proper forms in the output.

We segmented both sides of the corpus into morphemes based on the analyses, so the tokens in our system were morphemes instead of word forms. The following is a lowercased example sentence pair from the training corpus:

a[det] 137[szn] apró[mn] csillag[fn] [ela] álló[mn] spirál[fn] meg+[ik] duplázódik[ige] [me3] .[punct]

the_dt spiral_nn of_in 137_cd tiny_jj star_nn s_nns double_vb ed_vbd itself_prp ._.

The motivation for this approach was that Hungarian has a very rich morphology with thousands of possible inflected forms for each word in the open word classes. In addition, many English function words, such as prepositions, possessive and other pronouns etc. correspond to bound morphemes in Hungarian, which makes already the word alignment part of the Moses training procedure a difficult task. It is difficult capture generalizations like the ones above using a word form based representation. There are also systematic morpheme order differences between these corresponding morphemes: the inflectional suffixes (or postpositions) corresponding to English prepositions follow noun phrases rather than preceding them and the same applies to possessive pronouns and subject pronouns (the latter corresponding to verb agreement suffixes). We hoped that these difficulties could be addressed by a morpheme based solution adequately.

The phrase table was built using the default *grow-diag-final* heuristic from *Giza++* alignments that we acquired from the morpheme based representation of the corpus. We used the default settings for Giza++. We also used a lexicalized reordering table. The distortion parameter was left at the default value. We also analyzed and tried to use a 5-gram language model built from the monolingual English corpus that was published as part of the WMT09 shared translation task training material but the resulting model was too big to be loaded into the 3GB RAM of the machine that we used in our experiments. We tried to use IRSTLM instead of SRILM but we did not manage to solve the memory overload problem. So in the end we used a 5-gram morpheme based language model that was built from the English side of the bilingual training corpus only.

We run the MERT parameter optimization procedure using a morpheme based BLEU score computed on the morpheme segmented version of the WMT09 Hungarian to English tuning set. MERT took several days to run.

4.2 Results

We used the parameter settings suggested by the (morpheme BLEU score based) MERT optimization and generated English surface word forms using *morphg*. We expected that the morpheme based solution would pose a new problem: that of misplaced morphemes in the output that do not correspond to any valid surface word form. In such cases we resorted to skipping the misplaced morpheme, although this is obviously not an optimal solution.

The BLEU score we obtained on the detokenized output was not very encouraging, to put it mildly: 7.82. When we rerun the decoder with the parameter settings obtained from a previous broken down MERT session, we obtained somewhat better results: 7.95. But this is still very far from the 9.96/10.10 points achieved by MetaMorpho and the hybrid solution. Inspection of the translation results confirmed that the translations generated by the morpheme based setup are far inferior to those generated by our rule based system.

Inspecting Giza++ alignments revealed that, contrary to our hopes, segmenting the training corpus into morphemes did not in itself solve the word alignment quality problem: the alignments look even worse than those achieved on the plain text version of the corpus. On the other hand, all the drawbacks of the approach that we predicted: reduced span of local dependencies in the language models and the phase table due to the increased number of tokens spanning the same span of input, misplaced morphemes, etc. seem to have hit us.

5 Conclusion

In this article, we described the rule based, hybrid and statistical systems that we implemented and used in the WMT09 shared translation task.

Although we only managed to slightly improve the performance of our rule based machine translation system in our hybrid experiment and

with our first attempt at a morpheme based statistical system we obtained more modest results than we hoped, we think that it is still worth to make further attempts to build better translation systems for the Hungarian English language pair along these lines.

Acknowledgments

References

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos – an open source trigram tagger In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions,* Association for Computational Linguistics, Prague, Czech Republic, 209–212.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions,* Association for Computational Linguistics, Prague, Czech Republic, 177–180.

Guido Minnen, John Carroll and Darren Pearce. 2001. *Applied morphological processing of English,* Natural Language Engineering, 7(3). 207–223.

Xuan-Hieu Phan. 2006. *CRFTagger: CRF English POS Tagger,* http://crftagger.sourceforge.net/

Gábor Prószéky and Attila Novák. 2005. Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.): *Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday,* 116–125. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford.

Gábor Prószéky and László Tihanyi. 2002. MetaMorpho: A Pattern-Based Machine Translation System. In: *Proceedings of the 24th 'Translating and the Computer' Conference,* 19–24. ASLIB, London, United Kingdom.